Modelos Probabilistas Aplicados

Johanna Bolaños Zúñiga

Matricula: 1883900

Tarea 7

1. Introducción

En el presente trabajo se analizan datos generados con diferentes funciones para mostrar la aplicación de la regresión lineal y el uso de las transformaciones. Este análisis es complementado con resultados obtenidos de experimentos computacionales realizados en el software R versión 4.0.2 [4]. El código empleado se encuentra en el repositorio GitHub [1].

2. Regresión lineal

El análisis de regresión se centra en la exploración, explicación y estudio de dependencia de una variable mediante una o más variables explicativas. El término regresión significa que utilizaremos información pasada y es lineal porqué está bajo el supuesto de que entre dos variables (x y y) existe una relación lineal. Cuando se utiliza sólo una variable independiente para tratar de explicar la variable dependiente, es una regresión lineal simple, pero cuando se utilizan más de dos variables independientes o dependientes, se conoce como regresión múltiple.

La regresión lineal consiste en generar un modelo de regresión (ecuación de una recta) que permita explicar la relación lineal que existe entre dos variables. A la variable dependientes o respuesta se le identifican como y, y a la variables predictoras o independientes como x. Las ecuaciones 1 y 2, son las ecuaciones estimadas de regresión lineal simple y múltiple, respectivamente,

$$y = b_1 x + b_0 + \epsilon, \tag{1}$$

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \epsilon, \tag{2}$$

donde b_0 es la ordenada al origen (el valor de y cuando x es igual a cero) y, b_1 es la pendiente de la recta (el cambio en y cuando x aumenta en una unidad y ϵ es el error aleatorio. Este último representa la diferencia entre el valor ajustado por la recta y el valor real. Recoge el efecto de todas aquellas variables que influyen en y pero que no se incluyen en el modelo como predictores. Al error aleatorio también se le conoce como residuo. Los errores, se consideran variables aleatorias independientes distribuidas normalmente con media cero y desviación estándar σ. Esto implica que el valor medio o valor esperado de $\hat{y}$ son los mostrados en las ecuaciones 3 y 4, las cuales son las ecuaciones estimadas de regresión lineal simple y múltiple, respectivamente,

$$\hat{y} = \hat{b}_1 x + \hat{b}_0 + \epsilon, \tag{3}$$

$$\hat{y} = \hat{b}_1 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_p + \epsilon, \tag{4}$$

donde $\hat{y}$ es el valor estimado (aproximado) de y, $\hat{b}_0$ y $\hat{b}_p$ son estimaciones que se conocen como coeficientes de regresión, ya que toman aquellos valores que minimizan la suma de cuadrados residuales, dando lugar a la recta que pasa más cerca de todos los puntos.

En conclusión, el análisis de regresión consiste en definir la variable independiente x que ayude a explicar (estimar) la variable dependiente y, siempre que exista una relación lineal entre ellas, además, de que ambas variables deben ser cuantitativas [2].

Para determinar si hay relación lineal entre las variables dependientes e independientes, por lo general, se utilizan las gráficas de dispersión, ya que son ayudas visuales que permiten observar, rápidamente, si existe esta relación. Sin embargo, un análisis más fuerte es determinar la correlación entre los datos, ya que es una medida de la presencia de una relación lineal en datos bivariados (entre dos variables). Hay diferentes definiciones, sin embargo, la comúnmente utilizada, es la de correlación de Pearson (se puede hallar con la función `cor(x, y)`). Se denota con la letra r y sus valores se interpretan de la siguiente manera:

- Cerca de uno: cuando x crece, y crece de manera linealmente dependiente.
- Cerca de menos uno: cuando x crece, y disminuye de manera linealmente dependiente (o vice versa).
- Cerca de cero: no está presente ninguna relación lineal entre x y y.

El valor de la correlación se puede hallar con la función `cor(x, y)` en el programa R. De igual manera, la función `lm(y~x)`, donde y, es la variable dependiente (la que se trata de predecir) y x es la variable predictora (independiente), permite obtener el modelo estimado de regresión lineal de los datos a analizar. Con la función de `summary` se generan los resultados de este modelo.

Para entender el funcionamiento de la función `lm` se realizó una serie de experimentos, en los cuales consideramos crear diversas funciones, tanto lineales como no lineales, para la obtención de datos dependientes (variables `y`).

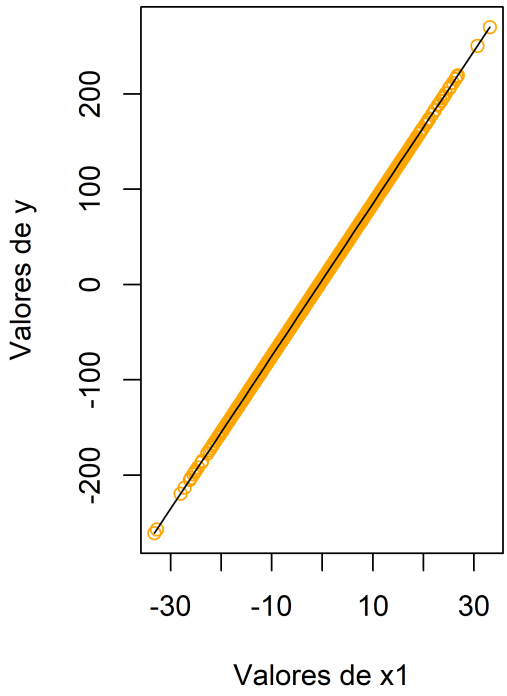

Figura 1: Diagrama de dispersión de los datos provenientes de la ecuación lineal

Para el ejemplo de la regresión lineal simple, se generaron $1,000$ datos a partir de la ecuación $y = 8x_1 + 5$, donde x_1 es un número pseudoaleatorio entre 10 y 60. Los resultados obtenidos fueron guardados en un dataframe. Se realizó un diagrama de dispersión (ver figura 7 para apreciar la relación entre las variables x_1 y y, en la cual se observa, claramente, la relación lineal entre estas variables.

test.txt

```
Call:
lm(formula = y ~ x1, data = f1)

Residuals:
       Min         1Q     Median         3Q        Max
-3.870e-14 -2.100e-14 -1.760e-14 -1.360e-14  1.241e-11

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 5.000e+00  3.532e-14 1.415e+14   <2e-16 ***
x1          8.000e+00  9.188e-16 8.707e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 4.281e-13 on 998 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.58e+31 on 1 and 998 DF,  p-value: < 2.2e-16
```

De acuerdo al resultado obtenido de la aplicación de la función `lm`, se observa que los coeficientes de regresión $\hat{b}_1 = 8$ y $\hat{b}_0 = 5$ y $r = 1$, es decir, cuando $x_1 = 0$, es valor estimado de $y = 8$ y por cada unidad que aumente x_1 el valor estimado de y aumenta en 8 unidades. Por lo tanto, el modelo estimado de regresión lineal para estos datos estaría dada como $\hat{y} = 8x_1 + 5$ (el cual con anterioridad se tenía). También se realizó el mismo procedimiento considerando que los valores de la variable x_1 tuvieran una distribución exponencial, normal o uniforme, y se obtuvieron los mismos resultados.

Para la regresión lineal múltiple, también se generaron $1,000$ datos a partir de la ecuación $y = 20 * x_1 + 5 * x_2 + 1$, donde las variable x_1y x_2 son pseudoaleatorias uniformes. Los resultados obtenidos fueron guardados en un dataframe. Se utilizo la función `cor` para determinar la correlación entre estas variables. Las respectivas correlaciones se encuentran graficadas en la figura 2, en la cual se puede observar la dependencia lineal de la varible y con las variables x_1 y x_2

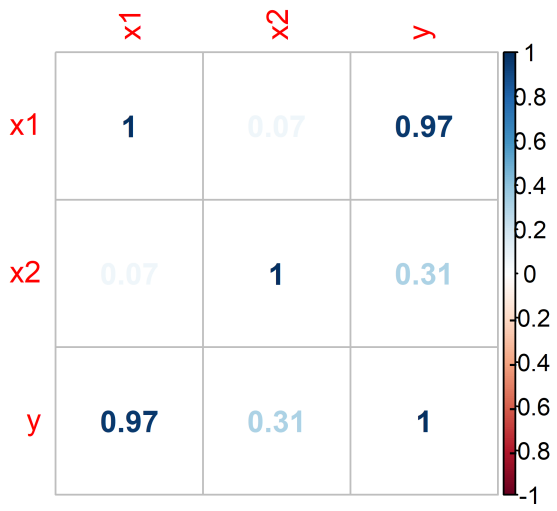

Figura 2: Correlación entre los datos provenientes de una ecuación lineal con dos variables independientes

test.txt

```
Call:
lm(formula = y ~ x1 + x2, data = f3)

Residuals:
```

```
       Min         1Q     Median         3Q        Max
-3.578e-13 -2.000e-16  3.200e-16  8.600e-16  2.560e-14

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 1.000e+00  9.223e-16 1.084e+15   <2e-16 ***
x1          2.000e+01  1.249e-15 1.601e+16   <2e-16 ***
x2          5.000e+00  1.270e-15 3.938e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.139e-14 on 997 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.401e+32 on 2 and 997 DF,  p-value: < 2.2e-16
```

De acuerdo al resultado obtenido de la aplicación de la función `lm`, se observa que $\hat{b}_1 = 20$, $\hat{b}_2 = 5$, $\hat{b}_0 = 1$ y $r = 1$. Por lo tanto, el modelo estimado de regresión lineal múltiple para estos datos estaría dada como $\hat{y} = 20x_1 + 5x_2 + 1$. También se realizó el mismo procedimiento considerando que los valores de las variables x_1 y x_2 tuvieran una distribución exponencial, normal o sin una distribución específica y se obtuvieron los mismos resultados.

Hasta aquí, se ha realizado el análisis con datos de los cuales se conocía su comportamiento, sin embargo, en los casos reales, este comportamiento no se conoce y no siempre es lineal, lo cual aumenta la complejidad del análisis de estos y la predicción de su comportamiento, haciendo necesario ajustar la tendencia de los datos a un modelo lineal ya que esto facilita el análisis y la predicción de su comportamiento para la toma de decisiones. Para llevar a cabo este ajuste se utilizan las transformaciones.

3. Transformaciones

La meta de las transformaciones es acomodar los datos a una relación lineal, ya que bajo este criterio se facilita el análisis y la predicción del comportamiento de estos.

De acuerdo a Mangiafico [3], para datos sesgados a la derecha o izquierda (sesgo positivo y negativo, respectivamente), las transformaciones comunes incluyen raíz cuadrada, raíz cúbica y logaritmo. Otro enfoque es utilizar la transformación de Box-Cox, la cual determina un valor lambda (λ), que se utiliza como coeficiente de potencia para transformar los valores. El procedimiento Box–Cox se realiza mediante la función `boxcox` en el programa R. Esta utiliza un procedimiento de *log-likelihood* para encontrar la lambda que se utilizará para transformar la variable dependiente de un modelo lineal (una regresión lineal). Cabe mencionar que estas transformaciones solo se pueden aplicar sobre variables de respuesta

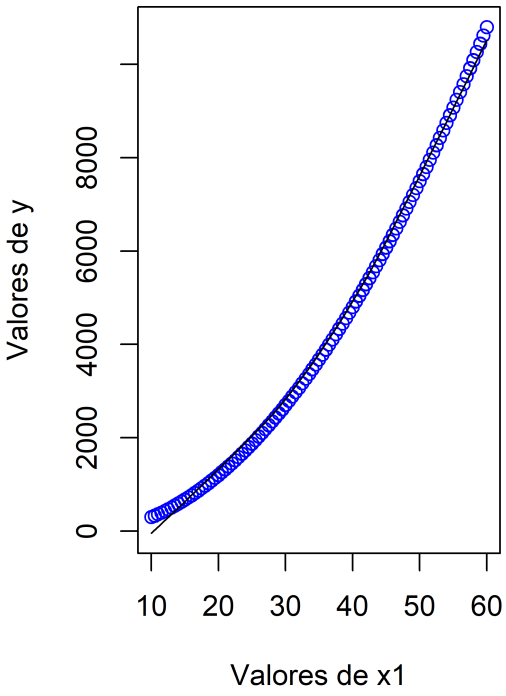

Figura 3: Diagrama de dispersión de los datos generados a partir de una función cuadrática

(variables dependientes) estrictamente positivas.

En este trabajo se utilizó las transformaciones logarítmicas, raíz cuadrada y Box–Cox. Para determinar que tipo de transformación es la más apropiada, se utilizó el coeficiente de determinación (R^2), el cual es una medida de la bondad de ajuste de la recta estimada a los datos reales, es decir, entre más cercano esté este valor a la unidad, la transformación tendrá un mejor ajuste. Este coeficiente se calcula elevando al cuadrado el valor de r en los modelos de regresión lineal simple el valor de (R^2), sin embargo, no es así en regresión múltiple. Existe una modificación de (R^2) conocida como ($R^2 - ajustado$) que se emplea principalmente en los modelos de regresión múltiple, el cual introduce una penalización cuantos más predictores se incorporan al modelo.

Para el análisis de las trasformaciones, se generaron $1,000$ datos a partir de la ecuación cuadrática de la forma $y = 3x_1 1^2 + rnorm(n)$, donde x_1 es un número pseudoaleatorio entre 10 y 60 y n son las réplicas. Los resultados obtenidos fueron guardados en un dataframe. Se realizó un diagrama de dispersión (ver figura 3 para apreciar la relación entre las variables x_1 y y), en la cual se observa, claramente, que hay una relación entre estas variables, pero no es lineal, por lo que se procedió a realizar tres diferentes transformaciones.

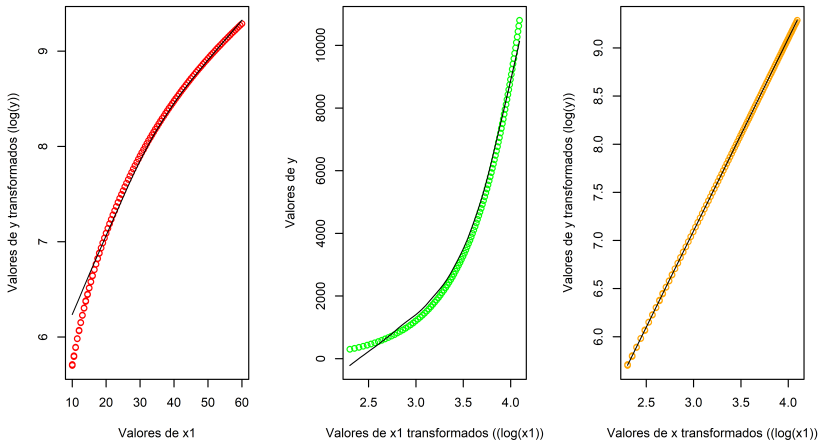

Figura 4: Comparativo de diagramas de dispersión con la transformación logarítmica

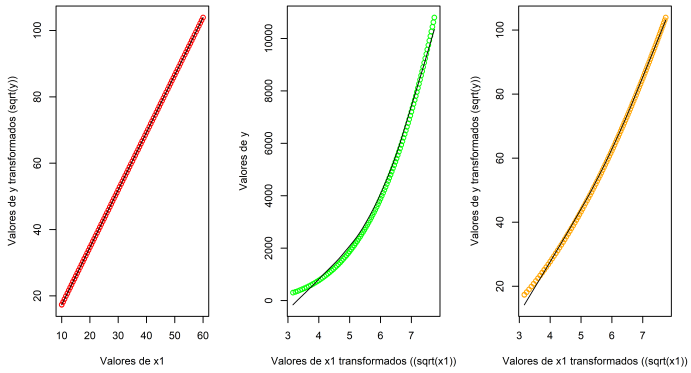

Figura 5: Comparativo de diagramas de dispersión con la transformación de raíz cuadrada

En la figura 4, se muestra el comparativo de los diagramas de dispersión de las transformaciones logarítmicas aplicadas, tanto a los datos de las variables x, como a y y a ambas, en la cual podemos observar que se obtiene un ajuste lineal a los datos cuando se aplica la transformación logarítmica a x y y.

En la figura 5, se muestra el comparativo de los diagramas de dispersión de las transformaciones de raíz cuadrada aplicadas, tanto a los datos de las variables x, como a y y a ambas, en la cual podemos observar que se obtiene un ajuste lineal a los datos cuando se aplica la transformación logarítmica a los datos de la variable, así mismo, su coeficiente de determinación $R^2 = 1$.

Finalmente, en la figura 6, se muestra el diagrama de dispersión de la transformación Box–Cox, en la cual se puede observar que al igual que con la transformación logarítmica de x y y o la raíz cuadrada de y, los datos se logran ajustar a relación lineal. Además, para estas trasformaciones el coeficiente de

determinación fue $R^2 = 1$. El valor de lambda para estos datos es de $\lambda = 0.505$.

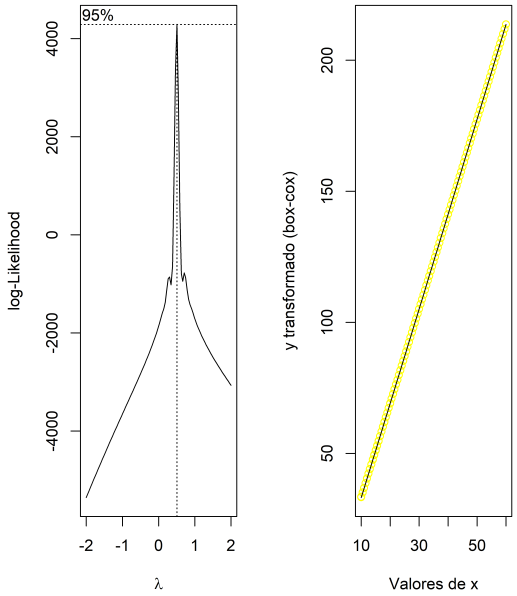

Figura 6: Valores de lambda (λ) y diagrama de dispersión con la transformación de Box–Cox

test.txt

```
Call:
lm(formula = (y^lambdabox - 1)/lambdabox ~ x1, data = f2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16687 -0.09227 -0.02531  0.07517  0.44237

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.074042   0.008967  -342.8   <2e-16 ***
x1           3.609200   0.000232 15559.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1065 on 998 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.421e+08 on 1 and 998 DF,  p-value: < 2.2e-16
```

Para efectos prácticos, se consideraron los datos obtenidos de la aplicación de la función `lm` con la transformación de Box–Cox (aunque también es válido escoger cualquiera de las tres buenas trasformaciones obtenidas), para determinar el modelo estimado de regresión lineal simple. De acuerdo a los

resultados arrojados, se observa que $\hat{b}_1 = 3.60$, $\hat{b}_0 = -3.07$. Por lo tanto, el modelo estimado de regresión lineal simple para estos datos estaría dado como $\hat{y} = 3.60x_1 - 3.07$. Adicionalmente, se realizó el mismo procedimiento considerando que los valores de la variable x_1 tuvieran otro tipo de distribución y para este caso, no se obtuvieron los mismos resultados, haciendo que las transformaciones propuestas no fueron suficientes. En el programa R, la rutina `assumptions(lm)` de la librería *trafo*, permite obtener información acerca de varias transformaciones que cumplan con los supuestos de un modelo lineal de los datos que estamos analizando.

Para la regresión lineal múltiple, también se generaron $1,000$ datos a partir de la ecuación no lineal con dos variables independientes. Los resultados obtenidos fueron guardados en un dataframe. Se utilizo la función `cor` para determinar la correlación entre estas variables. Las respectivas correlaciones se encuentran graficadas en la figura 2, en la cual se puede observar que no hay una dependencia lineal de y con las variables x_1 y x_2

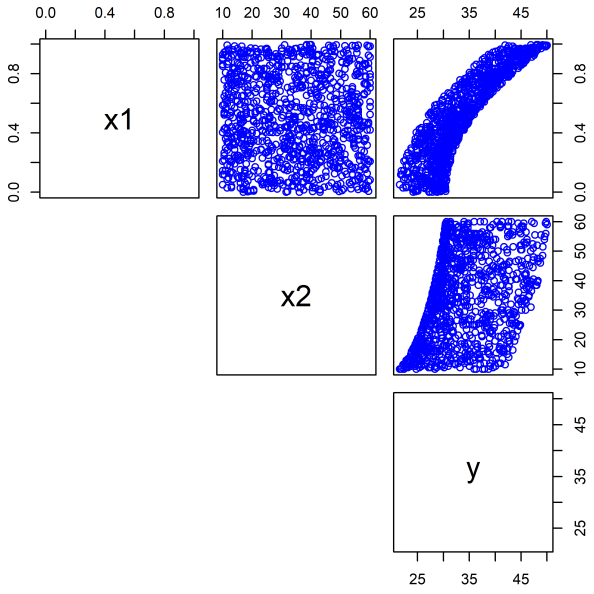

Figura 7: Correlación entre los datos provenientes de una ecuación no lineal con dos variables independientes

Se aplicaron las transformaciones logarítmicas, de raíz cuadrada y Box–Cox, de las cuales, la transformación $(\log y)$ presentó el mejor $R^2 = 0.9519$, es decir, es capaz de explicar el $95.19\,\%$ de la variabilidad observada en los datos. De acuerdo a los resultados arrojados, se observa que $\hat{b}_1 = 0.586$, $\hat{b}_2 = 0.005$ y $\hat{b}_0 = 3.035$. Por lo tanto, el modelo estimado de regresión lineal múltiple para estos datos estaría dado como $\hat{y} = 0.586x_1 + 0.005x_2 + 3.035$, lo que significa que, si el resto de variables se mantienen constantes, por cada unidad que aumenta el predictor en cuestión, la variable y varía en promedio tantas unidades como indica la pendiente. Para este ejemplo, por cada unidad que aumenta el predictor x_1, el

valor de y aumenta en promedio 0.586 unidades, manteniéndose constante la variable x_2.

test.txt

```
Call:
lm(formula = log(y) ~ x1 + x2, data = f4)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11258 -0.02824 -0.00467  0.02808  0.11054

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.035e+00  4.049e-03  749.50   <2e-16 ***
x1          5.863e-01  4.517e-03  129.81   <2e-16 ***
x2          5.089e-03  8.849e-05   57.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04128 on 997 degrees of freedom
Multiple R-squared:  0.9519,    Adjusted R-squared:  0.9518
F-statistic:  9861 on 2 and 997 DF,  p-value: < 2.2e-16
```

Referencias

[1] Bolaños Z., Johanna. Repositorio en GitHub de la clase de modelos probabilistas aplicados. Recursos libre, disponible en github.com/JohannaBZ/Probabilidad/tree/master/Tarea7, 2020.

[2] Gutiérrez B. Ana. *Probabilidad y Estadística, Enfoque por competencias.* McGraw Hill, 2012.

[3] Mangiafico, Salvatore S. Summary and Analysis of Extension Program Evaluation in R. Recurso disponible en, https://rcompanion.org/handbook/I_12.html.

[4] The R Foundation. The R Project for Statistical Computing. https://www.r-project.org/, 2020.